様式８

論　文　内　容　要　旨

報告番号	甲　先　第　436　号	氏　名	Robert Nsinga
学位論文題目	USING COMPUTING FIRST PRINCIPLES TO IMPROVE THE SYMBIOTIC PERFORMANCE IN ALGORITHMS AND PROCESSORS USED IN LOW-POWERED MACHINE LEARNING （コンピューティングの第一原理を使用して、低電力の機械学習で使用されるアルゴリズムとプロセッサの共生パフォーマンスを向上させる）		

内容要旨

Using less electric power or speeding up processing is catching the interests of researchers in deep learning. Models have grown in complexity and size using as much precision depth as can be computationally supported regardless of how expensive the minimum required cooling system might cost. Quantization has offered ease of deployment to small devices lacking floating precision capability, but little has been suggested about the floating numbers themselves. This thesis evaluates hardware acceleration for embedded devices that cannot support the energy requirements of floating numbers and proposes solutions to challenge the limits of power consumption and apply them to measure their effectiveness in terms of energy demand and speed capacity.

Experts have declared the end of Moore's law with the current state of nanotechnology coming to terms with its inability to increase the performance per transistor density ratio. Accelerators, although providing a countering measure, have also increased their power needs to unsustainable levels. At the same time there has been sufficient increase in knowledge, such as distributed computing, to branch-off into possibilities that could reduce power demands while maintaining, or possibly increase microprocessor performance. This thesis highlights some important challenges that were born out of the rapid rise of deep learning.

We present experimental results showing that low-powered devices can serve as powerful tools in low cost deep learning research. In doing so we are interested in slowing down the ongoing trend that favors expensive investment in deep learning computers. Using known properties in computer architecture, hardware acceleration, and digital arithmetic we implement ways to design algorithms that symbiotically match their performance in accordance with the theoretical limits afforded by the hardware components that run them.

Computer processors are utilized based on their ability to execute instructions defined in code or machine-readable format. Some processors are multi-purpose, others are domain-specific, the former being good at a wide range of tasks and the latter only focused for specific tasks. While executing any task an ideal processor should engage all its transistors to ensure that no part is left underutilized. However, in practice it is not always the case, which is why domain-specific processors are optimized to carry only the instructions for which they would fully commit their components.

It is considered good practice when algorithms are designed to encourage the maximum use of available capacity for any execution. Our proposed method improves the symbiotic complementarity in peak algorithm performance and theoretical hardware capacity.

Our evaluation shows gains in throughput and latency on a few common algorithms used in deep learning. We show up to 21% gain in energy savings and up to 22% reduced latency while maintaining optimal output error (2.1% average relative error). Our measurements are based on multiply-accumulate (MAC) with throughput, or performance expressed in the number of MACs per second, and energy in pico-Joules per MAC. Processor manufacturers publish theoretical numbers for speed, usually expressed in FLOPS (floating operations per second) which we based our estimated latency reduction on.

Assistive intelligence is a growing field provides interfaces where continuous input is expected throughout execution. The foundations of computing architecture can be revisited without stalling the progress achieved. Hybrid hardware-software tinkering is open to improvement and better numerical approximations.